MSc Technical Medicine – M3 Thesis

Final report

Project C.R.A.N.I.U.M.
(Constructing a Real-time Alarm for Nearing Intracranial hypertension Using Machine learning)

Thesis:
Final Report

Student:
Sybren van Hal (4383842)

Supervisors:
M: M. van der Jagt
T: J. F. Veenland

Date:
26-05-2021

Universiteit Leiden
The Netherlands

TU Delft Delft University of Technology

Erasmus University Rotterdam

Universiteit Leiden

TU Delft Delft University of Technology

ERASMUS UNIVERSITEIT ROTTERDAM

PROJECT C.R.A.N.I.U.M.

Constructing a Real-time Alarm for Nearing Intracranial hypertension Using Machine learning

Sybren-Willem Theodorus van Hal

Student number : 4383842

26-05-2021

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

University Leiden ; Delft University of Technology ; Erasmus University Rotterdam

Master thesis project (TM30004 ; 35 ECTS)

Intensive Care Unit, Erasmus MC

November 2020 – May 2021

Supervisors:

Dr. Jifke Veenland

Radiology & Medical Informatics

Dr. Mathieu van der Jagt

Intensive Care Unit

Thesis committee members:

Prof. dr. ir. Jaap Harlaar, TU Delft (chair)

Dr. Mathieu van der Jagt, Erasmus MC

Dr. Jifke Veenland, Erasmus MC

Prof. dr. Diederik Gommers, Erasmus MC

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus Universiteit Rotterdam

Abstract

Introduction

Intracranial hypertension (IH) is a harbinger of secondary brain injury in patients suffering from traumatic brain injury (TBI), can be mitigated at the Intensive Care Unit (ICU) and is associated with a poor prognosis. Current clinical practice consists of treating IH once it has occurred, by medical or surgical interventions. This is later than desired, as secondary injury has already been initiated. A pre-emptive approach may be preferable, and seems possible since many physiological variables that may aggravate IH are known and can be managed clinically. The aim of this research is to develop a machine learning method that is able to predict whether or not a patient will develop IH in the near future during ICU stay.

Methods

A cohort of 114 patients with TBI admitted to the ICU of Erasmus MC was selected. Long Short Term Memory (LSTM) models were trained and evaluated with 26 clinical variables to predict IH. The effect of the length of the minimal IH period, the length of the prediction window and the number of included variables was evaluated. Primary outcome measures were the model loss, accuracy, and Area Under the receiver operating characteristic Curve (AUC).

Results

We achieved a mean AUC of 0,83 [95% CI: 0,68-0,98] with a model predicting periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes and using only the ICP and mean arterial blood pressure (MAP). All models showed decreasing training and validation loss values during the first few epochs of model training. Thereafter, the training loss continued to decrease while the validation loss started to increase.

Conclusion

We developed a LSTM model that was able to predict, with a mean AUC of 0.83 [95% CI: 0,68-0,98], the occurrence of IH after half an hour based on the ICP and MAP. Adding more clinical variables resulted in overtrained models.

Introduction

Physiological situation

Intracranial hypertension (IH) is an augury of secondary brain injury in patients suffering from traumatic brain injury (TBI), and is associated with a poor prognosis[1]. The initial trauma causes hematomas or contusions, often accompanied by tissue edema. These are all space-occupying lesions. The Monro-Kellie doctrine states that the sum of the volumes of the brain tissue, cerebrospinal fluid (CSF), and intracranial blood, is constant[2]. If another volume, such as a hematoma, is introduced in the cranium, the contents of the cranium are compliant to a certain extent, because the amount of intracranial CSF may be reduced. However, increasing occupation of space eventually leads to an increase in intracranial pressure (ICP). IH is present if the ICP exceeds a certain threshold (generally 20-25 mmHg[3]). This hampers the cerebral perfusion, leading to brain ischemia, which is secondary harm on top of the initial impact of TBI.

Clinical situation

Patients suffering from severe TBI, often with reduced levels of consciousness, are admitted to the Intensive Care Unit (ICU), where they are carefully monitored. Among many other parameters, the ICP is continuously measured. Current clinical practice consists of treating IH once it has occurred, by medical or surgical interventions[4]. This is later than desired, as secondary injury has already been initiated. A pre-emptive approach may be preferable, and seems possible since many physiological variables that may aggravate IH are known and can be managed clinically. If healthcare professionals would be able to foresee an imminent IH event, preventive measures may be taken such as mitigation of factors that are known to contribute to increases in ICP.

Proposed improvement

Some established clinical variables that are known to contribute to an increased ICP include fever, hypo-osmolality of serum (inducing cerebral edema), hyperglycemia, prolonged hyperventilation or hypoventilation, venous congestion caused by high PEEP levels of the ventilator, and fluid overload[5]. These factors are represented by variables that are continuously measured at the intensive care unit (e.g. temperature, glucose, sodium, partial pressure of carbon dioxide, partial pressure of oxygen, and blood pressure) and are highly amenable to treatment. Furthermore, all TBI patients receive at least one CT-scan of the brain in order to evaluate the intracranial damage. These measured variables and CT-scan(s) may contain valuable information that might enable the prediction of an IH event[6]. Utilizing machine learning, we may be able to harness this information to predict IH in real-time.

Aim of this research

This pilot study is part of a master's thesis for the MSc Technical Medicine. The aim of this research is to train a machine learning model that is able to predict whether or not a patient is going to develop IH in the near future during ICU stay.

Methods

Literature study

We performed a systematic review on the prediction of IH in patients with TBI using artificial intelligence. This literature study is enclosed in Appendix 1. Based on the outcome of our literature review and existing knowledge of physiological variables that are associated with IH[5], we created a list of 21 variables that we suspected to be interesting to use for the development of a machine learning model. A list of these variables is available in Appendix 2.

Data collection

We asked the Erasmus MC department of Data & Analytics for all recorded measurements of the clinically measured variables we deemed of interest, of 30 patients that had been admitted to the ICU of the Erasmus MC in 2018, and had also been enrolled in an earlier study. This process required the development of a research protocol, a Data Protection Impact Assessment, and a Standard Operating Procedure regarding the safeguarding of patient data. These documents are available as supplementary material. We received data of 30 patients, consisting of 12 variables. A full list containing these variables is provided in Appendix 3.

During this project, a PhD candidate at the Erasmus MC ICU was working on the development of a patient dashboard, for which he developed a method to extract data from the electronic health records. We asked him for all recorded measurements of the clinically measured variables we deemed of interest, of 100 patients that had been admitted to the ICU of the Erasmus MC in 2019-2021. We received data of 114 patients, consisting of 28 variables. A full list containing these variables is provided in Appendix 3.

We asked the Erasmus MC Imaging Trial Bureau for the CT-scans of the head, corresponding to the patients in the data query for the Erasmus MC department of Data & Analytics. Besides a research protocol, this process required a statement by a Medical Research Ethics Committee that this research is not subject to the Medical Research Involving Human Subjects Act. We submitted our research protocol to the Medical Research Ethics Committee, but we have not yet received a response. Thus, we have not been able to include CT-scans in this research.

All used data for this project has been anonymized, and was analyzed as such. The key file representing the link between the anonymized data and the identity of the patients is stored at a secured Erasmus MC computer and can only be accessed by one of the researchers, who is an intensivist and has legal access to the patient data because of the professional standards (“beroepsgeheim”).

Data processing

We used Anaconda (version 4.9.2) to develop a custom Python (version 3.7.6.final.0) script, utilizing Keras (version 2.4.3) running on TensorFlow (version 2.3.1) for the development of a machine learning model. The goal of this script was automatic processing of patient data, drawing samples, and training a machine learning model to predict whether or not a data sample precedes IH. We used the dataset of 30 patients for development of the script. A summary of the capabilities of this script may be found in Appendix 4. The script, including extensive explanatory comments, and a list of all used packages, are available as supplementary material.

After construction of the Python script, we used the dataset of 114 patients, as this dataset contained the most patients and variables. Due to time restrictions we were unable to include the fluid input and output data, hence in total, the data of 26 variables was used. The median age of the patients in this dataset was 47,5 (Interquartile range: 31-64), with 71% being male.

For every variable, we took the mean value every five minutes. This was done to reduce the dimensionality of the data and to reduce the impact of missing values and outliers. Remaining missing values were replaced using linear interpolation. Starting from the first measurement, the first three hours of data were discarded, as during these first hours, a patient is undergoing several diagnostic and sometimes acute therapeutic procedures which may yield unstable ICP values. Therefore, we focused on the period beyond the first three hours when most patients have entered a more stable phase. We scaled the data per variable such that the median equals zero, and that the first quartile and third quartile of the interquartile range equal minus one and one, respectively. We did this to standardize the values that will eventually be analyzed by the machine learning model. We performed automatic detection of periods of IH using a prespecified ICP threshold and prespecified minimum time above the threshold.

A data sample length of one hour was chosen, based on the findings in the literature review. This sample length does not change throughout this research. Two types of samples were collected: data samples preceding IH and data samples not preceding IH. Data samples preceding IH were collected a prespecified amount of time in advance of the IH period. This prespecified amount of time is called the “prediction window”. The minimal IH period length was also prespecified. At most, three data samples preceding IH were collected per patient. In figure 1, a visual representation of a theoretical sample preceding IH is provided. Note that a data sample consists of an hour of ICP data that does not meet the prespecified properties of intracranial hypertension, along with the data of other variables during that time.

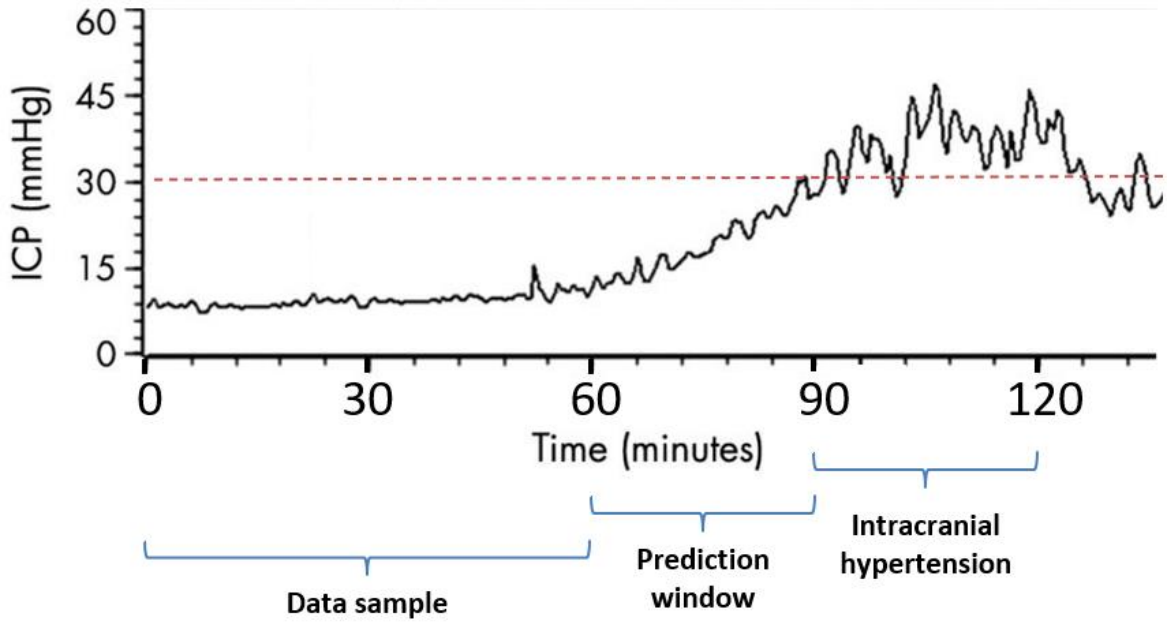

Figure 1. Visual representation of a data sample preceding a period of intracranial hypertension. We automatically detect periods of intracranial hypertension, and sample an hour of patient data that is located a “prediction window” away. In this case, the threshold for intracranial hypertension equals ICP≥30mmHg, and the prediction window equals 30 minutes. Adapted from Czosnyka M et al[7].

Data samples not preceding IH were randomly collected from all patients. Such a data sample was only collected if no IH period was present in both the theoretical prediction window and the minimum IH period length following the sample. At most, three data samples not preceding IH were collected per patient.

The data samples were randomly split in a training set (60%), validation set (20%) and test set (20%). Samples of one type (preceding IH or not preceding IH), originating from a single patient remained together in one of the three sets. For patients that experienced IH, it was

possible that samples preceding IH could be placed in one of the three data sets, while the samples not preceding IH could be located in another one of the three data sets.

The type of patient data that was used in this research was time series data, which implies a series of chronologically ordered data points. We decided to create a Long Short Term Memory (LSTM) machine learning model for data analysis, because this type of model is especially suitable for classifying time series data[8]. We chose to make a model consisting of two layers, with eight neurons per layer. Each model was trained 10 times, in order to obtain the average performance. A model was always trained for 50 epochs with a batch size of eight.

Experiments

During training the LSTM model aims to learn certain patterns in the data that allows it to create decision rules, which eventually enables the model to decide whether or not a sample is likely to precede IH, or not. During the learning process, the model takes a “batch” of random samples from the training set. We prespecified the size of this batch to be eight samples. The model then tries to create decision rules based on the samples in the batch. Subsequently, the model tries these rules on all samples in the validation set; acting as a binary classifier that calculates the likelihood of a sample belonging to one of two classes. The amount of incorrect predictions is represented by the “loss” value. We prespecified the loss value to be calculated using binary cross-entropy, which is a popular loss function for binary classifiers[9]. In essence, this function introduces a penalty score for bad predictions, which is called the “loss”. The function calculates the mean of all losses that are obtained by trying the decision rules on the samples in the validation. Thus, if many samples in the validation set are predicted correctly, the mean loss will be low, whereas a lot of wrong predictions result in a higher mean loss. The LSTM model tries to optimize its decision rules by minimizing the loss. After processing a batch, the decision rules of the model are updated based on the loss. Once the model has processed all batches in the training set, an “epoch” has passed. The model may then be improved further by going through this process again, with the samples in the training set being shuffled across the batches. We prespecified the amount of epochs to be 50.

We performed three experiments. First, we trained models using similar IH definitions and identical prediction windows as found in our systematic review, and using the data of 26 variables. Thus, we trained a model to predict periods of ICP≥30mmHg lasting at least 10 minutes, with a prediction window of 30, and a model to predict periods of ICP ≥20mmHg lasting at least 15 minutes, with a prediction window of 30.

Second, we investigated 30-minute, one-hour, three-hour, and six-hour prediction windows, in combination with two IH definitions: ICP≥20mmHg lasting at least 15 minutes, and ICP≥20mmHg lasting at least 30 minutes. We used the data of 26 variables for these models. We chose a threshold of 20 mmHg for these models, as the Erasmus MC protocol regarding ICP management after neurotrauma strives to keep the ICP below 20mmHg[10].

Third, we trained a model using a similar IH definition, an identical prediction window and the same variables as the best performing model in our systematic review. Thus we trained a model to predict periods of ICP ≥20mmHg lasting at least 15 minutes, with a prediction window of 30 minutes, using only the ICP and mean arterial blood pressure (MAP).

The primary outcome measures of this research were the model loss, accuracy, which represents the number of correct predictions divided by the total number of predictions[11], and de Area Under the receiver operating characteristic Curve (AUC), which may be interpreted as the chance that the model, when given a random sample preceding IH and a random sample not preceding IH, ranks the former sample higher (in terms of “possibly preceding IH”) than the latter[12]. The AUC was obtained by using a model on the test data.

Results

We were unable to calculate sensitivity and specificity values. For every experiment, we visualized the loss values per epoch of all models, for all 10 times each model was trained. We did the same for the accuracy values per epoch, and the eventual receiver operating characteristic (ROC) curve. These figures are provided in Appendix 5.

In Appendix 5, figure 1, an interpretational example is provided on how to read and understand a figure containing the loss values obtained during training.

During experiment 1, we achieved a mean AUC of 0,72 and a mean accuracy of 0,87 with the model prediction periods of ICP≥30mmHg lasting at least 10 minutes, using a prediction window of 30 minutes. We obtained a mean AUC of 0,85 and a mean accuracy of 0,86 with the model predicting periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes. Table 1 provides an overview of these results The figures containing the loss and accuracy values during training, and the ROC curves obtained by testing the models on test samples, are provided in Appendix 5, figures 2-4.

The mean AUC and mean accuracy values of the models trained during experiment 2 are provided in Table 2. The best mean AUC was achieved by the model predicting periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 3 hours. The figures containing the loss and accuracy values during training, and the ROC curves obtained by testing the models of experiment 2 on test samples, are provided in Appendix 5, figures 5-7.

The model trained during experiment 3, predicting periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes and using only the ICP and MAP, achieved a mean AUC of 0,83 and a mean accuracy of 0,91. The figures containing the loss and accuracy values during training, and the ROC curves obtained by testing the model on test samples, are provided in Appendix 5, figures 8-10.

For all experiments, the figures containing the loss values obtained during training of the models showed decreasing training and validation loss values during the first few epochs. Thereafter, the training loss continued to decrease while the validation loss started to increase. This phenomenon is called “overfitting”, and is explained in Appendix 5, figure 2.

Table 1. Model information and performance from experiment 1. Two models were trained using similar IH definitions and identical prediction windows as found in our systematic review, and using the data of 26 variables. An “event” is a sample preceding IH.

Prediction window	IH definition	Training samples (events)	Validation samples (events)	Test samples (events)	Mean AUC [95% CI]	Mean Accuracy [95% CI]
30 minutes	ICP≥30m mHg for at least 10 minutes	228 (27)	77 (8)	80 (8)	0,72 [0,62-0,82]	0,87 [0,85-0,89]
30 minutes	ICP≥20m mHg for at least 15 minutes	235 (31)	80 (11)	83 (12)	0,85 [0,81-0,89]	0,86 [0,84-0,88]

Table 2. Model information and performance from experiment 2. Models were trained using 30-minute, one-hour, three-hour, and six-hour prediction windows, in combination with two IH definitions: ICP≥20mmHg lasting at least 15 minutes, and ICP≥20mmHg lasting at least 30 minutes. The data of 26 variables was used. An "event" is a sample preceding IH.

Prediction window	IH definition	Training samples (events)	Validation samples (events)	Test samples (events)	Mean AUC [95% CI]	Mean Accuracy [95% CI]
30 minutes	ICP≥20m mHg for at least 15 minutes	235 (31)	80 (11)	83 (12)	0,85 [0,81-0,89]	0,86 [0,84-0,88]
30 minutes	ICP≥20m mHg for at least 30 minutes	232 (28)	76 (7)	73 (4)	0,84 [0,76-0,92]	0,93 [0,92-0,94]
1 hour	ICP≥20m mHg for at least 15 minutes	233 (29)	81 (12)	83 (14)	0,79 [0,72-0,86]	0,84 [0,83-0,85]
1 hour	ICP≥20m mHg for at least 30 minutes	228 (24)	76 (7)	76 (7)	0,74 [0,66-0,82]	0,88 [0,86-0,90]
3 hours	ICP≥20m mHg for at least 15 minutes	228 (24)	83 (14)	80 (11)	0,88 [0,85-0,91]	0,87 [0,86-0,88]
3 hours	ICP≥20m mHg for at least 30 minutes	228 (24)	74 (5)	76 (7)	0,75 [0,70-0,80]	0,91 [0,90-0,92]
6 hours	ICP≥20m mHg for at least 15 minutes	222 (18)	81 (12)	79 (11)	0,70 [0,63-0,77]	0,85 [0,83-0,87]
6 hours	ICP≥20m mHg for at least 30 minutes	225 (21)	73 (4)	74 (5)	0,75 [0,70-0,80]	0,92 [0,90-0,94]

Table 2. Model information and performance from experiment 3. A model was trained using a similar IH definition, an identical prediction window and the same variables (the intracranial pressure and mean arterial blood pressure) as the best performing model in our systematic review. An "event" is a sample preceding IH.

Prediction window	IH definition	Training samples (events)	Validation samples (events)	Test samples (events)	Mean AUC [95% CI]	Mean Accuracy [95% CI]
30 minutes	ICP≥20m mHg for at least 15 minutes	238 (34)	81 (12)	81 (12)	0,83 [0,68-0,98]	0,91 [0,88-0,94]

Discussion

Main findings

We were able to train a LSTM model predicting periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes and using only the ICP and MAP, that achieved a mean AUC of 0,83 and a mean accuracy of 0,91. This is in concordance with the literature we found in the systematic review[13-15].
Using 26 variables to train models resulted in overtraining. The amount of positive cases was small: for every model, about 5-15% of all samples preceded IH. In order to train models using more variables than the ICP and MAP effectively, more positive samples are needed. The average AUC appears to decrease as the prediction window length increases. However, this should be investigated using a larger dataset.

Limitations

The Python script may still be improved. We propose the following considerations.

After calculation of the mean value for every variable every 5 minutes, remaining missing values are filled using linear interpolation. Missing values after the last known measurement of a variable are interpolated as well using forward filling, while missing values before the first measurement of a variable remain unaffected. This may possibly be solved with backward filling using the value of the first known measurement.

Samples not preceding IH that were collected from a patient that experienced IH, may end up in a different set (i.e. training set, validation set or test set) than samples preceding IH, collected from the same patient. We do not suspect that this influences the model, although this may be changed so that both types of sample are placed in the same set. This ensures that the model is not able to train an validate using samples originating from the same patient.

The data concerning daily fluid input and output was not used in this research due to time limitations. Implementing the use of this additional data in training a LSTM model may improve its results.

The script is not yet able to read CT scans and utilize features from those. This ability should be added as soon as CT scans are available, so that LSTM models may be trained with imaging features.

The script is able to detect periods of ICP above a certain threshold lasting a certain amount of time. Sometimes, short periods of ICP above the threshold follow shortly after one another; perhaps these multiple short periods should be treated as one, if the time between them is short enough.

The impact of outliers is reduced by taking the mean value of a variable every 5 minutes. An additional way of mitigating outlier influence could be the implementation of a minimum and maximum threshold level for each variable, and ignoring any values that exceed those thresholds.

The mean value of a variable is calculated every 5 minutes, followed by linear interpolation to fill any remaining missing values. For variables that are not measured frequently, this means that at many time points, an interpolated value is used. This could result in cases where a data sample only contains interpolated values for a certain variable. Such a situation cannot occur in clinical practice, because if the most recent data of a patient would be analyzed, it is impossible to perform linear interpolation using future measurements. A possible solution would be increasing the measurements frequency, but since that is not always achievable in practice, forward filling using the last measured values may be an option.

Samples not preceding IH seem over-represented, as the part of samples preceding IH was only 5-15% in the data sets of every experiment. This is not surprising because samples not preceding are collected from every patient, while samples preceding IH may only be obtained from the patients that experienced IH at least once. In addition, the maximum amount of samples preceding IH could not always be collected due to unavailability. It may be interesting to investigate the performance of a model that was trained on data that contained roughly equal amounts of both sample types. This may be achieved by taking a small portion ("subsampling") of the negative samples.

The scaling of the data per variable is currently performed before splitting the data into the training, validation and test sets. However, the scaling should actually happen after the data has been split. Not doing so has likely biased the model evaluation, as information may have leaked from the test set to the training set[16].

The scaling of the data has also led to skewed values, as some columns contain many zeros. This is probably caused by wrong measurements followed by linear interpolation between zeros or, if the last measured value was a zero, by forward filling by the interpolation function. These zeros are still taken into account when determining the median and interquartile range, leading to a distorted scaling process. This may be solved by implementing a lower and upper threshold for variable values, as mentioned earlier in this section.

It would be interesting to see which variables are most important for the model. If we can figure out what variables possess the most predictive values, we could try to train a model using only those variables, leaving out any redundant variables that may make the model unnecessarily complicated.

We used automatically detected periods of IH. In practice, a high ICP may have various causes, such as transportation of the patient. In order to be sure that automatically predicted samples precede clinically relevant IH, the medical records of a small group of patients could be searched for periods of IH for which treatment was given. The information from the records could be checked against the automatically detected periods, to investigate whether or not automatic detection of IH captures all or most of the clinically relevant IH periods, without detecting too much periods that are not clinically relevant.

Future research

It is evident that models trained using 26 variables, suffer from overtraining. There are several possibilities to solve overtraining, such as by decreasing the amount of layers in the model, decreasing the amount of neurons, or by scaling the data[17]. Future research could focus on this. Furthermore, we discuss the limitations of this research and provide possible solutions in the previous section, which could be taken into account in future research.

Implementation

Implementation of a machine learning model in clinical practice comes with various practical challenges and ethical considerations.

One of the first challenges finds its origin in patient data. The model needs quick access to the latest relevant patient data, in order to make real-time predictions. This may require integration of the model in the current hospital systems. Furthermore, when the machine learning model has processed data of a patient, it may be possible to add this data to a database that may be used to periodically update the prediction model. This should be done in accordance with the General Data Protection Regulation (GDPR). Patients should provide informed consent before any of their data is saved in such a database, after discharge. Should the patient pass away during their admission, relatives of the patient may provide informed consent.

Furthermore, implementation may require the medical staff to have some basic understanding of machine learning models and how to interpret their results. Especially nurses should receive some training, as they are most likely to notice periods of IH. However, this does not guarantee a problem-free implementation and use. Currently, decisions of a machine learning model may never be fully understood, due to the intrinsic mathematical complexity behind such a decision[18]. Also, the model provides a decision along with a confidence score, usually ranging from zero (no confidence) to one (absolutely sure). This confidence score differs from the confidence score a human may give its own decision[19]. This complicates situations where the output of a model differs from a clinician's opinion: comparing both confidence levels is nearly impossible, and the reason behind the decision of the model cannot be deciphered. Since a machine learning model is usually trained using data of hundreds of patients, which includes the medical expertise of a similar amount of clinicians, a model will never be 100% accurate[20]. It may be regarded as a new source of information, but also as a source of uncertainty for clinicians. Adequate training of healthcare professionals in the use of these prediction tools and protocol development is of key importance.

Another ethical question arises when a healthcare professionals thinks that a patient is stable, while the model warns that the patient is probably developing intracranial hypertension. Should immediate action be undertaken, such as administering medication to the patient, or would it be better to exert a wait-and-see policy? A follow-up question could be: what are the consequences when the medical professionals choose to do nothing, and the patient indeed develops IH, eventually leading to permanent disability? These questions should not lead to fear among clinicians to chose to ignore a model's outcome. Furthermore, clinicians should be wary of possible overtreatment when a model's outcome is different from their own views. Following the computer could lead to unnecessary administration of medication or interventions, resulting in preventable harm for the patient, and the wasting of resources such as materials, time, and money. Adequate training and protocol development is needed before such tools can be put into practice.

Should clinicians choose to adhere to the computer's prediction, it may very well be impossible to check if the right choice has been made. That is: if medication that lowers the ICP is given pre-emptively to a patient, the patient may not develop intracranial hypertension. How do we know whether or not the patient would have developed IH if no action was undertaken? It may very well never be possible to determine if a patient was overtreated. Studying these situations would be arguably impossible, as it would be unethical to withhold therapy for one patient, while a similar patient does receive treatment.

Conclusion

We trained a LSTM model based on the ICP and MAP that was able to predict the occurrence of periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes. This model achieved a mean AUC of 0,83. Adding more physiological variables resulted in overtrained models. To solve this, future research could focus on improvements in the data processing, simplifying the model, finding out which variables are most important. Furthermore, longer prediction windows could be investigated.

References

1 Pinto VL et al. Increased Intracranial Pressure [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan [updated 2020 Jul 20; cited 2021 May 12]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK482119/

2 Mokri B. The Monro-Kellie hypothesis: applications in CSF volume depletion. Neurology. 2001 Jun 26; 56(12): 1746-8

3 Nourallah B et al. Critical thresholds for intracranial pressure vary over time in non-craniectomised traumatic brain injury patients. Acta Neurochir (Wien). 2018; 160(7): 1315–1324

4 Schizodimos T et al. An overview of management of intracranial hypertension in the intensive care unit. J Anesth. 2020 May 21: 1–17

5 Sharma S et al. Intracranial Hypertension [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan [updated 2020 May 23; cited 2020 Oct 15]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK507811/

6 Fernando SM et al. Diagnosis of elevated intracranial pressure in critically ill adults: systematic review and meta-analysis. BMJ. 2019; 366: l4225

7 Czosnyka M et al. Monitoring and interpretation of intracranial pressure. J Neurol Neurosurg Psychiatry. 2004 Jun; 75(6): 813–821

8 Hochreiter S et al. Long short-term memory. Neural Comput. 1997 Nov 15; 9(8): 1735-80

9 Nichols Ja et al. Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev. 2019 Feb; 11(1): 111–118

10 Van der Jagt M. ICP behandeling bij neurotrauma [Internet]. [cited 2021 May 20]. Available from: https://icv-erasmusmc.nl/protocol/icp-behandeling-bij-neurotrauma-2/

11 Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978 Oct;8(4):283-98

12 Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27(8), 861–874

13 Carra G et al. Prediction model for intracranial hypertension demonstrates robust performance during external validation on the CENTER-TBI dataset. Intensive Care Med. 2020 Oct 1

14 Güiza F et al. Novel methods to predict increased intracranial pressure during intensive care and long-term neurologic outcome after traumatic brain injury: development and validation in a multicenter dataset. Crit Care Med. 2013 Feb; 41(2): 554-64

15 Güiza F et al. Early Detection of Increased Intracranial Pressure Episodes in Traumatic Brain Injury: External Validation in an Adult and in a Pediatric Cohort. Crit Care Med. 2017 Mar; 45(3): e316-e320

16 Scikit-learn developers. sklearn.preprocessing.robust_scale [internet]. [cited 2021 May 18]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.robust_scale.html#sklearn.preprocessing.robust_scale

17 Yamashita R et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018 Aug; 9(4): 611-629

18 Burrell J. How the machine ‘thinks’: understanding opacity in machine learning algorithms. Big Data Soc 2016; 3(1)

19 Gigerenzer G, Hoffrage U, Kleinbölting H. Probabilistic mental models: a Brunswikian theory of confidence. Psychol Rev 1991;98(4):506–28

20 Grote T et al. On the ethics of algorithmic decision-making in healthcare. J Med Ethics. 2020 Mar; 46(3): 205–211

21 Brownlee J. How to use Learning Curves to Diagnose Machine Learning Model Performance [internet]. [updated 2019 February 27; cited 2021 May 25]. Available from: https://machinelearningmastery.com/learning-curves-for-diagnosing-machine-learning-model-performance/

Appendix

Appendix 1: literature review

Predicting intracranial hypertension in patients with traumatic brain injury using artificial intelligence: a systematic review

S-W. T. van Hal, BSc
M. van der Jagt, MD, PhD
J.F. Veenland, MSc, PhD

ABSTRACT

Introduction: Intracranial hypertension (IH) may lead to secondary injuries in patients suffering from traumatic brain injury (TBI). Current clinical practice consists of monitoring the intracranial pressure (ICP) and starting treatment once IH has been diagnosed. A pre-emptive approach may be more beneficial for patients. Previous studies have shown that predicting future IH events is possible. In this systematic review, we assess the available literature covering the prediction of ICP/IH using machine learning (ML). We aim to identify the used models and variables, and the resulting performance.

Methods: We searched the Embase and Ovid electronic databases for studies using ML to predict ICP/IH in TBI patients. We only included studies that performed internal or external validation. Article quality was determined using a custom quality assessment. We summarized the patient demographics, data characteristics, used models and variables, and performance measures.

Results: We retrieved 1934 non-duplicate publications, of which five were eligible for inclusion in this systematic review. In total, we identified five variables used for model development: ICP, mean arterial pressure (MAP), brain tissue oxygenation, pressure reactivity index, and time since last crisis. We found that the most commonly used predicting method consisted of a Gaussian Processes model utilizing the ICP and MAP. This appeared to be the best-performing prediction model to date, with a maximum reported area under the receiver operating characteristic curve of 0,93 when evaluated in an external dataset. Only two studies carried out external validation. All studies were retrospective; no studies described prospective, clinical use of ICP/IH prediction.

Conclusion: Research regarding ICP/IH prediction using ML is sparse. Some well-performing models have already been developed, but there is potential for improvement. Current literature reports the ability to predict an increase in ICP 60 minutes in advance, or an IH event 30 minutes in advance. In order to be more clinically relevant, earlier predictions are needed. ML-based ICP/IH predicting remains a promising concept to prevent secondary injury in TBI patients.

INTRODUCTION

Intracranial hypertension (IH) portends a worse prognosis in patients with traumatic brain injury (TBI) and should be treated expediently[1]. The primary brain injury, consisting of hematomas or contusions, is often accompanied by tissue edema resulting in IH. IH can be considered a harbinger of secondary injury, as it hampers cerebral perfusion and thus induces brain ischaemia[1]. Current practice has focused mainly on mitigating intracranial pressure (ICP) once it has occurred, by medical or surgical interventions[2]. However, current practice consists of applying countermeasures once the ICP has become too high (generally defined as exceeding 20-25 mmHg[3]). A pre-emptive approach that may contribute to prevention of ICP surges, by mitigation of contributing factors known to be able to induce secondary brain injury in patients at high risk for IH, may be preferable. Indeed, once the ICP rises, this is a clear sign of exhausted compensatory intracranial reserve (compliance) contributing to secondary injuries that would better be prevented instead of treated.

Established clinical variables contributing to secondary brain injury and IH include fever, hypo-osmolality of serum (inducing cerebral edema), hyperglycemia, prolonged hyperventilation or hypoventilation and venous congestion caused by high PEEP levels of the ventilator and fluid overload[4]. These factors are represented by variables that are continuously measured at the intensive care unit (e.g. temperature, glucose, sodium, pCO_2, pO_2, and blood pressure) and are highly amenable to treatment. Early identification of the risk of impending IH can enable clinicians or nurses at the bedside to correct derangement of these variables, alone or in combination, and thereby theoretically initiate countermeasures and decrease the risk of IH.

To date, prediction of IH with such physiological variables remains understudied. Furthermore, imaging results (e.g. Computed Tomography (CT) scans) may also harbor predictive features for IH, but studies are virtually absent[5]. Combining both the physiological and imaging features may result in an even higher potential to predict IH. Contemporary machine learning (ML) algorithms parallel or even outperform humans when, for example, predicting certain medical conditions[6] or analyzing images[7]. Hence, ML may be a valuable tool to help the identification of patients at risk of developing IH and guide preventive measures, such as strict avoidance of fever, or avoiding fluid overload[4].

The aim of this systematic review was to assess the available literature regarding the prediction of ICP/IH in TBI patients using ML, and subsequently answer the following questions: 1) What kind of ML methods and variables are being used to predict IH?; 2) What ML method yields the best results?; 3) Has prediction of IH with ML been validated externally, and is it already being used prospectively in clinical practice?

METHODS

The protocol for this systematic review has been registered in PROSPERO (registration number: CRD42020214744). This research was conducted and reported using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [8].

Search strategy

We searched both the Embase and Ovid electronic databases on 27-10-2020, for publications describing studies that involved intracranial pressure or intracranial hypertension, traumatic brain injury, and machine learning. The full queries can be found in **table 1**.

Study selection

The titles and abstracts of the retrieved studies were assessed by two authors (SvH, JV). Articles were included if they used machine learning with the aim to predict ICP/IH (and/or decreased cerebral perfusion pressure) in patients suffering from TBI, and reported performance measures on an internal or external validation set. We excluded articles not in English, articles without available full text, and duplicate articles. Based on the full text of the remaining articles, we identified and included all studies that met the inclusion criteria.

Data extraction

Data was extracted from the selected studies by one of the authors (SvH), using the following predefined list of items: population description, data specifics of training set, data specifics of validation set, variables used in model, sample frequency, prediction window length, IH definition, data cleaning process, type of validation, type of ML method used, Area Under the receiver operating characteristic Curve (AUC), accuracy (i.e. the number of correct predictions divided by the total number of predictions), sensitivity, and specificity. Among the aforementioned data specifics, the data instance length was collected. The term ”instance” is used to indicate a collection of data acquired during a certain time period (e.g. one hour), and has a label that show whether an IH event occurs during this period. These instances are subsequently used to teach a ML algorithm to recognize differences between data without any IH events, and data containing an IH event. When in doubt, authors discussed until consensus was reached.

As there are currently no reporting guidelines for ML prediction models available, we assessed the quality of each article using a custom quality assessment. This quality assessment can be found in **table 2** and is based on literature regarding the Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD)[9] and reporting guidelines of machine learning articles[10-12].

Main outcome

The main outcomes of this systematic review were the types of ML methods and variables used, prediction window length, and their resulting performance measures.

RESULTS

Study selection

We searched the Embase and Ovid electronic databases on 27-10-2020. In total, our search provided 1934 unique records, of which five[13-17] (four articles and one letter) were eligible for inclusion in this systematic review. A flowchart visualizing the article selection process is provided in **figure 1**.

Study characteristics

Four[13, 15-17] out of five studies mentioned the demographics of their cohort. Güiza F et al. (2017)[16] studied an adult and a pediatric cohort, and supplied demographics for both. The demographic features that were reported are summarized in **table 3.** The lowest and highest age in the reported interquartile ranges were 7,5[16] and 65[16], respectively. The median total Glasgow Coma Score (GCS) was six or seven for every study population, excluding the study by Myers RB et al. (2016)[17] that only reported the eye and motor GCS.

Table 4 summarizes the main study characteristics.

Data instance length ranged from 30 minutes[17] to four hours[15, 16].

To predict ICP/IH, all articles used preceding ICP, and four[13-16] out of five studies also used the mean arterial pressure (MAP). In addition to these two variables, one study[14] also used the brain tissue oxygenation and pressure reactivity index. One study[17] used only the ICP and the time since last crisis.

Data samples were taken every five seconds[14] up to every 72 seconds[17].

Data cleaning was described by three studies and consisted of removing values registered during an intervention[14]; removing obvious artifacts[16]; excluding physiologically impossible values[17]; interpolating missing data points[17]; and using a smoothing filter[17].

Nine different models were used. Güiza F et al. (2013)[15], Güiza F et al. (2017)[16] and Carra G et al. (2020)[13] used the same Gaussian Processes (GP) model, on different datasets. The models were able to predict the occurrence of IH 30 minutes in advance, and an elevation of the ICP 60 minutes in advance.

Table 5 shows the AUC, accuracy, sensitivity and specificity for each study where this was described. All publications provided at least the AUC. The article by Güiza F et al. (2017)[16] mentioned performance measures of the model for both the adult cohort and the pediatric cohort. This was also the only publication that provided 95% confidence intervals. Thus, in total, there were six AUC values, of which the average was 0,83. The best performance (an AUC of 0,93) was achieved by by Carra G et al. (2020)[13]. There were four[13-16] articles that mentioned an accuracy value, ranging from 61,5%[14] to 88%[13].Three[13, 15, 16] articles also reported the sensitivity and specificity values, ranging from 70%[16] to91[16] and from 48%[16] to 91%[13], respectively.

Four[13, 15-17] out of five studies investigated the prediction of IH specifically, while one[14] study only looked at the ICP course independent of a specific threshold. We found two different definitions of IH: ICP>30mmHg for 10 minutes[13, 15, 16], and ICP>20mmHg for 15 minutes[17]. The prediction windows varied from 30 minutes[13, 15-17] to one hour[14].

Three[14, 15, 17] studies performed internal validation and two[13, 16] performed external validation. All studies were retrospective; we did not find research that used ML in a prospective, clinical setting.

None of the included studies reported to have made their data or models publicly available.

Quality analysis

The quality scores for each publication is can be found in table 6. The average score was 50%, ranging from 33%[14] to 67%[16].

DISCUSSION

Main findings

In this systematic review on the utility of ML algorithms to predict ICP/IH in patients suffering from TBI, we found that only limited variables were consequently used (mainly ICP and MAP) and that prediction was limited to only 30 minutes in advance of the occurrence of IH, and 60 minutes in advance of a higher ICP. Gaussian Processes (GP) was most commonly used, followed by Logistic Regression.

The GP model by Güiza F et al. (2013)[15], based on the ICP and MAP, achieved the best AUC (0,93) when used on an external validation data set[13]. All studies were retrospective; we did not find literature describing prospective, clinical use of ML-based ICP/IH prediction.

Although the best performing model was developed on patient data from 2003-2005 (AUC: 0,872)[15] and was initially only internally validated, it still accomplished good results when validated externally on data from 2009-2013 (AUC: 0,90 (adult cohort) and 0,79 (pediatric cohort))[16] and 2015-2017 (AUC: 0,93)[13], indicating the robustness of the model.
Notably, the ML algorithms used mainly used pressure-related variables, with one study also using brain tissue oxygenation[14] and another study also using the time since last crisis[17].

The results we found raise the question whether or not ML will improve clinical practice as current models are focused on predictions 30 to 60 minutes in advance. It is arguable that this might offer enough advantage compared to clinical observation of an ICP trend directly preceding IH.

Limitations

This systematic review has several limitations. ICP/IH prediction with ML in patients with TBI is a niche topic; the available literature regarding this subject was scarce, with only four articles and one letter being eligible for inclusion. Two[13, 14] publications were of variable (<50%) quality. However, this quality assessment was custom-made, based on literature and what the authors deemed important in this type of studies. This makes the quality scores subjective. There are various articles concerning possible reporting guidelines for research involving a ML model, but there is a need for universal and widely acknowledged quality criteria.

Four[13, 15-17] out of five included studies described their population demographics. No study explicitly stated that all patient data from a specific time period were used, so we are unable to rule out cherry picking of patient data. Selection bias could therefore be a concern in selected studies for this review, since selecting only patients without missing data or artifacts and with very evident trends in the data may lead to a well-performing models, whereas using real-world data, might yield different prediction properties. Furthermore, no articles stated exactly how many data instances were sampled per patient, which hampers comparability of studies and insight into data collections underlying the machine learning algorithms.

Further, the included studies differed regarding the definition of IH, the used sample frequency and the used data instance length.

Furthermore, the lack of external validation of two models[12, 15] reduces the credibility of their reported results.

Future research

For future research, we can deduct several suggestions.

First, although a GP method produced the best results of all included studies, we cannot exclude that new types of ML methods, such as support vector machines and random forest, may have good or better performance when applied to variables that are used much earlier than the maximum of 60 minutes preceding IH. These methods are currently commonly used and should be considered for further research[18].
Second, mainly the ICP and MAP are utilized as variables in the included studies. It may be useful to take also homeostasis-related variables into account, although it is unknown whether these may further improve prediction. The ICP rise may be preceded by changes in homeostasis-related variables which can mostly be mitigated by clinical treatment, making them interesting from a clinical perspective.
Third, the use of imaging (especially CT-scan) features in the prediction of ICP/IH in patients with TBI holds promise in relation to ML. Future studies should explore the use of imaging features as (part of the) variables used to train a predicting method, since ML may especially be able to outperform human interpretation as it has been used before to automatically extract imaging features from CT-scans[19].
Fourth, we estimate the optimal required sample frequency for ML to be approximately a value per minute. The best performing model, by Güiza F et al. (2013)[15], also used this sample frequency. Utilizing a higher sample frequency results in many data points which may be challenging to analyze. On the other hand, lower sample frequencies may fail to capture changes early and may lead to skewed data, as one artifact could have drastic impact on a perceived trend.
Fifth, the training data instances varied from 30 minutes[17] to four hours[15, 16]. The necessary data instance length likely depends on the type of model, since the model by Myers RB et al. (2016)[17] used 30-minute instances and yielded an AUC of 0,86, whereas the best model by Feng M et al. (2012)[14], that also used this instance length, only achieved an AUC of 0,66. We suspect that an instance length of at least one hour should be used, to ensure that early warning signs can be picked up timely by a predicting model.
Sixth, the prediction window lengths varied from 30 minutes[13, 15-17] to one hour[14]. We suspect that a warning for imminent IH should be given at least 30 minutes prior to the projected event, to allow preventive measures to be taken. Güiza F et al. (2013)[15] affirmed the sufficiency of this prediction horizon. It may be even more clinically relevant if ML is able to identify patients, with high risk of developing IH during their entire stay, shortly after admission.
Seventh, In order to prevent selection bias as much as possible, an equal number of instances should be drawn from every patient, or at least a maximum amount of instances per patient.

CONCLUSION

There is a dearth of studies on ML-based prediction of IH in patients with TBI. Currently, the best performing method appears to be a GP model that utilizes the ICP and mean arterial pressure to predict IH within 30 minutes. New studies should consider using homeostasis-related variables and imaging features, possibly in combination with pressure-related variables.

REFERENCES

1 Pinto VL et al. Increased Intracranial Pressure [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan [updated 2020 Jul 20; cited 2020 Sept 9]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK482119/.

2 Schizodimos T et al. An overview of management of intracranial hypertension in the intensive care unit. J Anesth. 2020 May 21: 1–17.

3 Nourallah B et al. Critical thresholds for intracranial pressure vary over time in non-craniectomised traumatic brain injury patients. Acta Neurochir (Wien). 2018; 160(7): 1315–1324.

4 Sharma S et al. Intracranial Hypertension [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan [updated 2020 May 23; cited 2020 Oct 15]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK507811/.

5 Fernando SM et al. Diagnosis of elevated intracranial pressure in critically ill adults: systematic review and meta-analysis. BMJ. 2019; 366: l4225.

6 Flechet M et al. Machine learning versus physicians' prediction of acute kidney injury in critically ill adults: a prospective evaluation of the AKIpredictor. Crit Care. 2019 Aug 16; 23(1): 282.

7 Jammal AA et al. Human Versus Machine: Comparing a Deep Learning Algorithm to Human Gradings for Detecting Glaucoma on Fundus Photographs. Am J Ophthalmol. 2020 Mar; 211:
123-131.

8 Moher D et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev. 2015 Jan 1; 4(1): 1.

9 Moons KGM et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015 Jan 6; 162(1): W1-73.

10 Liu Y et al. How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. JAMA. 2019 Nov 12; 322(18): 1806-1816.

11 Luo W et al. Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016 Dec 16; 18(12): e323.

12 Stevens LM et al. Recommendations for Reporting Machine Learning Analyses in Clinical Research. Circ Cardiovasc Qual Outcomes. 2020 Oct; 13(10): e006556.

13 Carra G et al. Prediction model for intracranial hypertension demonstrates robust performance during external validation on the CENTER-TBI dataset. Intensive Care Med. 2020 Oct 1.

14 Feng M et al. Utilization of temporal information for intracranial pressure development trend forecasting in traumatic brain injury. Annu Int Conf IEEE Eng Med Biol Soc. 2012; 2012: 3930-4.

15 Güiza F et al. Novel methods to predict increased intracranial pressure during intensive care and long-term neurologic outcome after traumatic brain injury: development and validation in a multicenter dataset. Crit Care Med. 2013 Feb; 41(2): 554-64.

16 Güiza F et al. Early Detection of Increased Intracranial Pressure Episodes in Traumatic Brain Injury: External Validation in an Adult and in a Pediatric Cohort. Crit Care Med. 2017 Mar; 45(3): e316-e320.

17 Myers RB et al. Predicting Intracranial Pressure and Brain Tissue Oxygen Crises in Patients With Severe Traumatic Brain Injury. Crit Care Med. 2016 Sep; 44(9): 1754-61.

18 Shillan D et al. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Crit Care. 2019 Aug 22; 23(1): 284.
19 Wang S et al. Machine Learning and Radiology. Med Image Anal. 2012 Jul; 16(5): 933–951.

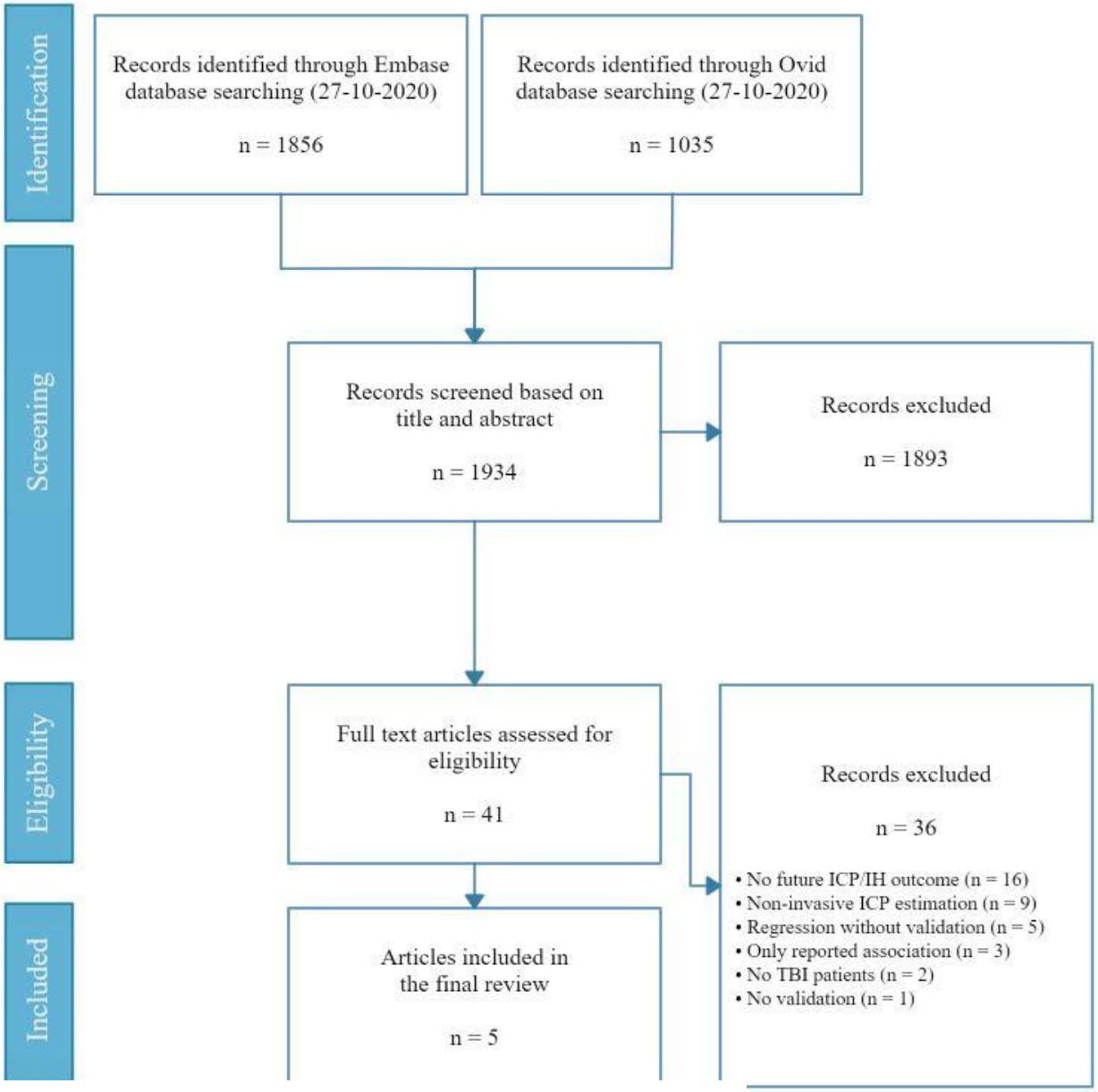

Figure 1: A flowchart representing the article selection process.

Database	Query
Embase	('intracranial hypertension'/exp OR 'intracranial pressure'/exp OR ('intracranial':ab,ti AND ('pressure':ab,ti OR 'hypertension':ab,ti)) OR 'ich':ab,ti) AND ('traumatic brain injury'/exp OR ('traumatic':ab,ti AND 'brain':ab,ti AND 'injury':ab,ti) OR 'tbi':ab,ti) AND ('machine learning'/exp OR (machine:ab,ti AND learning:ab,ti) OR model*:ab,ti OR computer*:ab,ti OR automat*:ab,ti OR neural*:ab,ti OR cnn:ab,ti OR predict*:ab,ti)
Ovid	('intracranial hypertension'/OR 'intracranial pressure'/OR ('intracranial'.ab,ti,kf AND ('pressure'.ab,ti,kf OR 'hypertension'.ab,ti,kf)) OR 'ich'.ab,ti,kf) AND ('traumatic brain injury'/OR ('traumatic'.ab,ti,kf AND 'brain'.ab,ti,kf AND 'injury'.ab,ti,kf) OR 'tbi'.ab,ti,kf) AND ('machine learning'/OR (machine.ab,ti,kf AND learning.ab,ti,kf) OR model*.ab,ti,kf OR computer*.ab,ti,kf OR automat*.ab,ti,kf OR neural*.ab,ti,kf OR cnn.ab,ti,kf OR predict*.ab,ti,kf)

Table 1: Search queries used in Embase and Ovid.

Criterion (points)	Min	Max
Patient demographics provided (+2)	0	2
No/minimal risk of bias due to patient selection (+3)	0	3
Amount of used data instances provided (+2)	0	2
Data instances are obtained equally from every patient (+2)	0	2
Variables used in model provided (+3)	0	3
Sample frequency provided (+3)	0	3
Prediction window is 30 minutes or more (+2)	0	2
Data cleaning process provided (+1)	0	1
Prospective study design (+2)	0	2
Validation set used: internal (+2) or external (+5)	0	5
Performance measures provided: sensitivity (+1), specificity (+1), AUC (+1), CI (+1)	0	4
Validated in clinical setting (+2)	0	2
Open science and data (+2)	0	2
Total	*0*	*33*

Table 2: Custom quality assessment for studies predicting future ICP or IH using machine learning. AUC = Area Under the receiver operating characteristic Curve. CI = Confidence Intervals.

Article	Patients in dataset	Age in years *Median (IQR) or Mean±SD*	Sex *% male*	GCS score *Median (IQR)*
Carra G et al. (2020)[11]	Training: N.A.	N.A.	N.A.	N.A.
	Validation: 257	47 (30-61)	81%	6 (3-10)
Feng M et al. (2012)[12]	Training: 82*	N.P.	N.P.	N.P.
	Validation: 82*	N.P.	N.P.	N.P.
Güiza F et al. (2013)[13]	Training: 178	33,1 (19-49)	80,9%	7 (4-10)
	Validation: 61	24 (13-44)	77,1%	7 (4-9)
Güiza F et al. (2017)[14]	Training: N.A.^{A}	N.A.^{A}	N.A.^{A}	N.A.^{A}
	Training: N.A.^{P}	N.A.^{P}	N.A.^{P}	N.A.^{P}
	Validation: 121^{A}	50 (28,5-65)^{A}	78%^{A}	7 (3-12)^{A}
	Validation: 79^{P}	10,4 (7,5-14,2)^{P}	74%^{P}	6 (5-8)^{P}
Myers RB et al. (2016)[15]	Training: 368	29 (21-40)	87%	7(4-9)**, 5 (2-5)***
	Validation: 261	30 (23-46)	85,1%	7 (3-8)**, 5 (2-5)***

*Table 3: Patient demographics of the training cohorts and validation cohorts. N = number of patients. GCS = Glasgow Coma Scale. IQR = interquartile range. SD = standard deviation. N.P. = not provided. N.A. = not applicable. ^{A}Adult cohort. ^{P}Pediatric cohort. *Same patients used for training and validation **Eye score. ***Motor score.*

Publication	ML methods	Variables	Outcome	Sampling	Training data	Validation data	Cleaning
Carra G et al. (2020)[11]	GP	Intracranial pressure Mean arterial pressure	IH (ICP>30mmHg for 10 minutes) 30 minutes in advance	N.P.	N.A.	External; N.P.	N.P.
Feng M et al. (2012)[12]	LogReg AODE AdaBoost-J48 BayesNet-K2 BayesNet-TAN LBR Naive Bayes SVM	Intracranial pressure Mean arterial pressure Brain tissue oxygenation Pressure reactivity index	ICP elevation, stability or reduction 60 minutes in advance	Value per 5 seconds	1-hour instances	Internal; 1-hour instances	Yes; used only data points between interventions
Güiza F et al. (2013)[13]	GP	Intracranial pressure Mean arterial pressure	IH (ICP>30mmHg for 10 minutes) 30 minutes in advance	Value per 60 seconds	2677 4-hour instances, 982 events (37%) from 108/178 patients (61%) *(patients with complete records)*	Internal; 1135 4-hour instances, 392 events (35%) from 33/61 patients (54%)	N.P.
Güiza F et al. (2017)[14]	GP	Intracranial pressure Mean arterial pressure	IH (ICP>30mmHg for 10 minutes) 30 minutes in advance	Value per 60 seconds	N.A.	External; 1051 4-hour instances, 231 events (22%) from 41/121 patients (34%)[A] 2219 instances, 811 events (37%) from 49/79 patients (62%)[P]	Yes; excluded obvious artifacts
Myers RB et al. (2016)[15]	GP LogReg AR-OR	Intracranial pressure Time since last crisis	IH (ICP>20mmHg for 15 minutes) 30 minutes in advance	Value per 72 seconds	43353 30-minute instances, 5979 events (14%) *(patients from 1989-1996)*	Internal; 38349 30-minute instances, 4025 events (10%) *(patients from 2006-2013)*	Yes; excluded physiologically impossible values, interpolated missing data, used smoothing filter

Table 4: This table summarises for each publication the used forecasting models and variables, the model outcome (i.e. ICP or IH as forecasting outcome, including IH threshold, and the prediction window), the used sample frequency, specifics regarding the data used to train and validate the models, and whether or not some sort of data cleaning has been performed, including a brief description. GP = Gaussian Processes. LogReg = Logistic Regression. AODE = Aggregating One-Dependence Estimators. AdaBoost-J48 = Ada-Boosting with Decision Tree. BayesNet-K2 = Bayesian Network with K2. BayesNet-TAN = Bayesian Network with TAN. LBR = Lazy Bayesian Rules. Naive Bayes = Naive Bayesian Classifier. SVM = Support Vector Machine. AR-OR = Autoregressive Ordinal-Regression. N.P. = not provided. N.A. = not applicable. [A] Adult cohort. [P] Pediatric cohort.

Article	Model	AUC	Accuracy	Sensitivity	Specificity
Carra G et al. (2020)[11]	GP	0,93	88%	83%	91%
Feng M et al. (2012)[12]	LogReg	0,645	62,1%	N.P.	N.P.
	AODE	0,66	62,4%	N.P.	N.P.
	AdaBoost-J48	0,632	61,5%	N.P.	N.P.
	BayesNet-K2	0,648	62,3%	N.P.	N.P.
	BayesNet-TAN	0,644	62,0%	N.P.	N.P.
	LBR	0,647	63,3%	N.P.	N.P.
	Naive Bayes	0,638	61,9%	N.P.	N.P.
	SVM	0,613	62,4%	N.P.	N.P.
	Average	*0,641*	*62,2%*	*N.A.*	*N.A.*
Güiza F et al. (2013)[13]	GP	0,872	77,4%	81,6%	75,2%
Güiza F et al. (2017)[14]	GP	0,90 [0,87–0,91][A]	86% [84–88][A]	70% [64-76][A]	90% [88–92][A]
	GP	0,79 [0,77–0,81][P]	64% [62–66][P]	91% [90–93][P]	48% [45–51][P]
Myers RB et al. (2016)[15]	GP	N.P.	N.P.	N.P.	N.P.
	LogReg	N.P.	N.P.	N.P.	N.P.
	AR-OR	0,86 [0,85-0,86]	N.P.	N.P.	N.P.
Average		*0,83*	*N.A.*	*N.A.*	*N.A.*

Table 5: Model performances. 95% confidence intervals are provided between brackets if they were reported. GP = Gaussian Processes. LogReg = Logistic Regression. AODE = Aggregating One-Dependence Estimators. AdaBoost-J48 = Ada-Boosting with Decision Tree. BayesNet-K2 = Bayesian Network with K2. BayesNet-TAN = Bayesian Network with TAN. LBR = Lazy Bayesian Rules. Naive Bayes = Naive Bayesian Classifier. SVM = Support Vector Machine. AR-OR = Autoregressive Ordinal-Regression. AUC = Area Under the receiver operating characteristic Curve. N.P. = not provided. N.A. = not applicable. [A] Adult cohort. [P] Pediatric cohort.

Article	Quality score
Carra G et al. (2020)[11]	15/33 (45%)
Feng M et al. (2012)[12]	11/33 (33%)
Güiza F et al. (2013)[13]	17/33 (52%)
Güiza F et al. (2017)[14]	22/33 (67%)
Myers RB et al. (2016)[15]	18/33 (55%)

Table 6: Quality scores.

Appendix 2: List of requested variables

We requested

- Intracranial pressure
- Mean arterial blood pressure
- Systolic blood pressure
- Brain tissue oxygenation
- Pressure reactivity index
- Shock index
- Pulse pressure
- Heart rate
- End tidal carbon dioxide
- Cerebral perfusion pressure
- Temperature
- Glucose
- Sodium
- Partial pressure of carbon dioxide
- Partial pressure of oxygen
- C-reactive protein
- Fluid balance per 24 hours
- Fluid intake per 24 hours
- Positive end expiratory pressure per 24 hours
- Ventilation tidal volume per 24 hours
- Ventilation peak pressure per 24 hours

Appendix 3: List of obtained variables

For 30 patients, we obtained data for the following variables:

- Intracranial pressure
- Heart rate
- End tidal carbon dioxide
- Temperature
- Glucose
- Sodium
- Partial pressure of carbon dioxide
- Partial pressure of oxygen
- C-reactive protein
- Positive end expiratory pressure
- Ventilation mean pressure
- Ventilation peak pressure

For 114 patients, we obtained data for the following variables:

- Intracranial pressure
- Non-invasive systolic blood pressure
- Non-invasive diastolic blood pressure
- Non-invasive mean blood pressure
- Systolic arterial blood pressure
- Diastolic arterial blood pressure
- Mean arterial blood pressure
- Pulse pressure variation
- Brain tissue oxygenation
- Heart rate
- End tidal carbon dioxide
- Cerebral perfusion pressure
- Temperature
- Glucose
- Sodium
- Partial pressure of carbon dioxide
- Partial pressure of oxygen
- C-reactive protein
- Fluid input
- Fluid output
- Positive end expiratory pressure
- Ventilation tidal volume
- Ventilation inspiratory tidal volume
- Ventilation expiratory tidal volume
- Ventilation tidal volume per body weight
- Ventilation peak pressure
- Ventilation plateau pressure
- Fraction of inspired oxygen

Appendix 4: Summary of the custom Python script

- Loading patient data provided in a Microsoft Excel file.
- Sorting the data per unique patient.
- Organizing the data in a structured format, while being able to detect measurements of multiple variables at a single point in time, creating a chronological collection of data with measurements or multiple variables per point in time.
- Calculate and present how much data has been collected per variable per patient.
- Calculating the means of all variables every prespecified time period and organizing these means in a structured format.
- Interpolating any missing mean values.
- Automatic detection of intracranial hypertension, based on a prespecified intracranial pressure threshold and prespecified time this threshold should be exceeded.
- Scaling of the means according to the interquartile range, making the values robust to outliers.
- Automatic random sampling of prespecified data periods preceding intracranial hypertension by a prespecified time.
- Automatic random sampling of prespecified data periods not preceding intracranial hypertension by a prespecified time, both in patients that did not experience IH and patients that did experience IH.
- Organizing the samples in a format that is suitable for training a machine learning model
- Training a machine learning model recognize periods of data preceding IH.
- Writing all output, including a logfile, in timestamped folder.

Appendix 5: Loss plots, accuracy plots and ROC curve plots

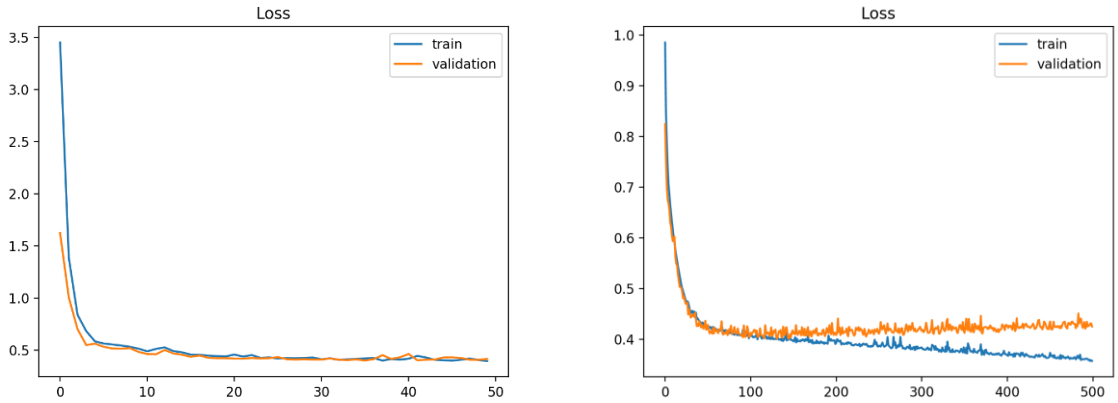

Figure 1. Interpretational examples of the model loss values during training. This figure on the left illustrates an ideal example, the figure on the right illustrates an example of overfitting. The loss values obtained by using the model on the training data and validation data, both decrease as the epochs pass. Eventually both lines should approximately overlap, indicating the ability of the model to predict classes in the training set with the same performance as in the validation set, as van be seen in the left figure. In case of "overfitting", the loss values obtained by using the model on the validation data, will increase while the loss values obtained by using the model on the training data remain low or decrease, as can be seen in the right figure. This means that the model learns to perfectly distinguish classes in the training data, while creating decision rules that are increasingly less general. This means the model may still perform well on the data in the validation and test set, as can be seen in the figures containing the accuracy values obtained by using the model on the validation data, or in the figures containing the Receiver Operator Characteristics curve obtained by using the model on the test data. However, a model that suffers from overfitting, is very likely to not perform well on a new dataset. Figures adapted from Brownlee J[21].

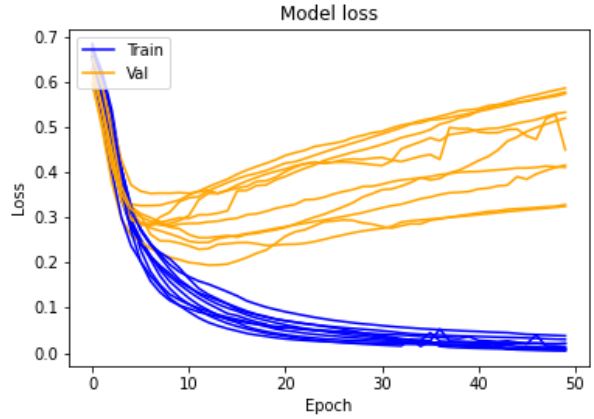

Figure 2A. Loss values per epoch during training of the 10 models aiming to predict periods of ICP≥30mmHg lasting at least 10 minutes, using a prediction window of 30 minutes.

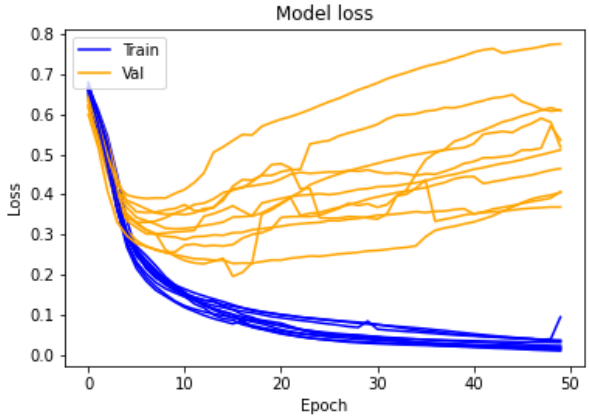

Figure 2B. Loss values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes.

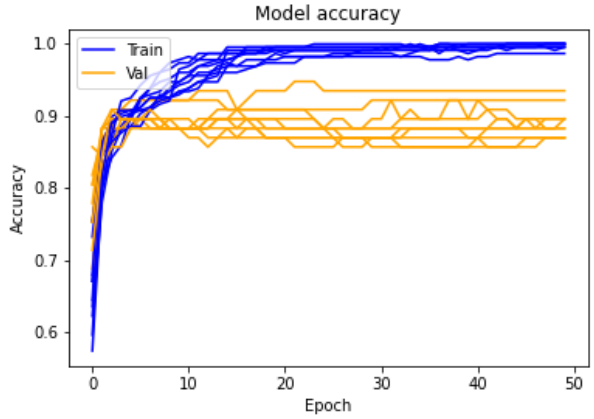

Figure 3A. Accuracy values per epoch during training of the 10 models aiming to predict periods of ICP≥30mmHg lasting at least 10 minutes, using a prediction window of 30 minutes.

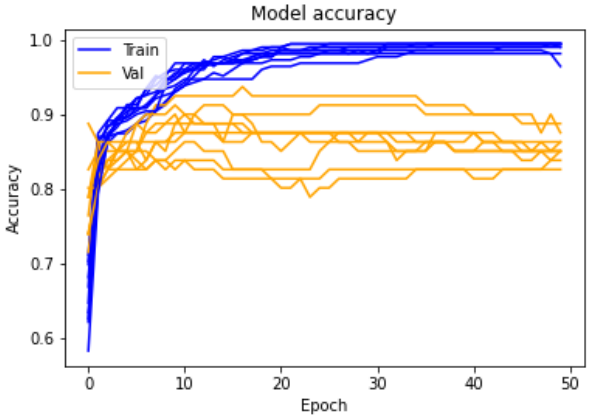

Figure 3B. Accuracy values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes.

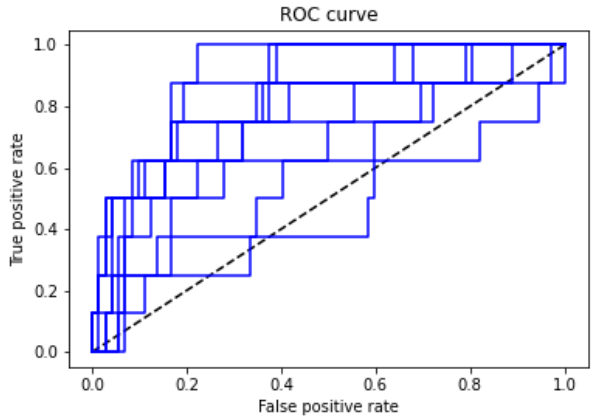

Figure 4A. Receiver Operator Characteristics curve of the 10 models aiming to predict periods of ICP≥30mmHg lasting at least 10 minutes, using a prediction window of 30 minutes, on test data.

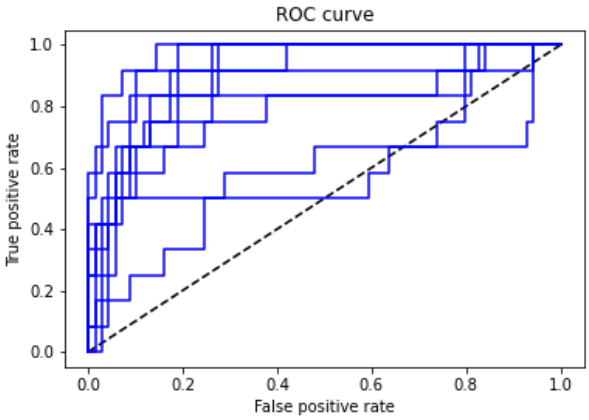

Figure 4B. Receiver Operator Characteristics curve of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes, on test data.

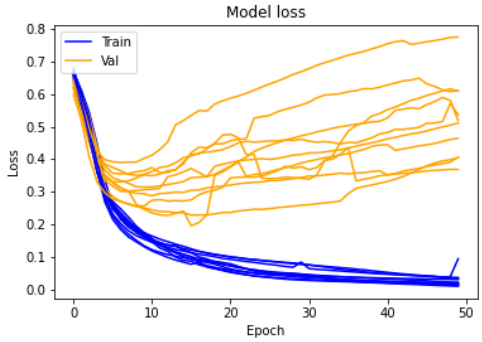

Figure 5A. Loss values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes.

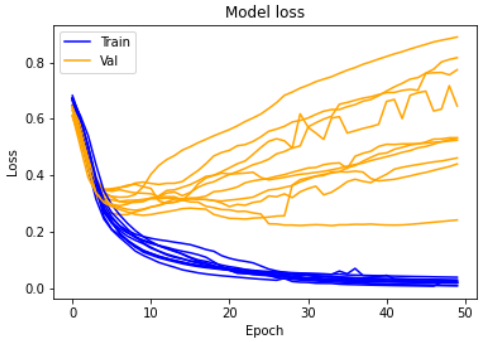

Figure 5B. Loss values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 30 minutes.

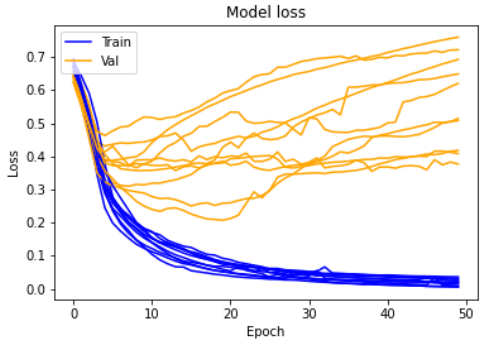

Figure 5C. Loss values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 1 hour.

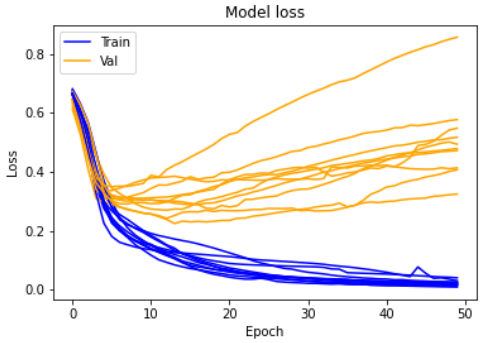

Figure 5D. Loss values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 1 hour.

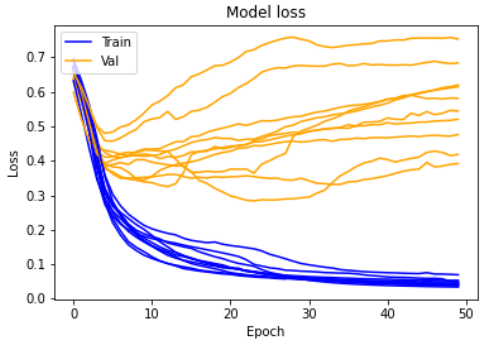

Figure 5E. Loss values per epoch during training set of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 3 hours.

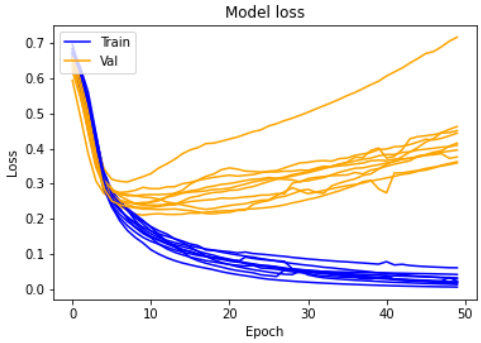

Figure 5F. Loss values per epoch during training set of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 3 hours.

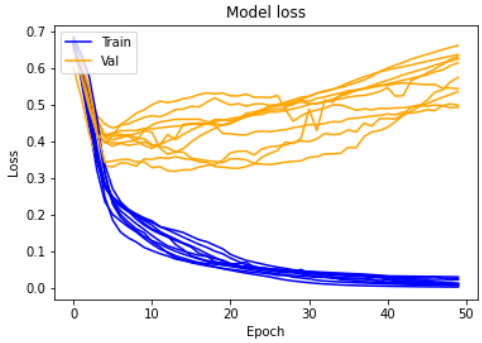

Figure 5G. Loss values per epoch during training set of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 6 hours.

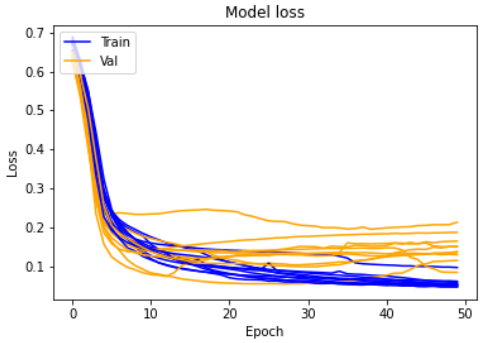

Figure 5H. Loss values per epoch during training set of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 6 hours.

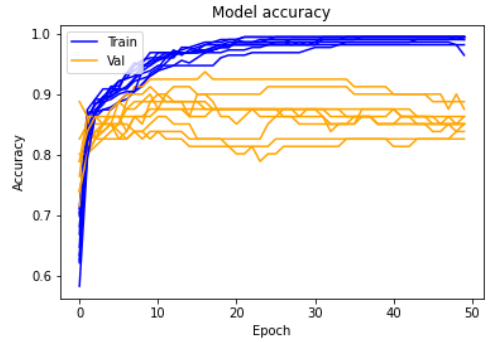

Figure 6A. Accuracy values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes.

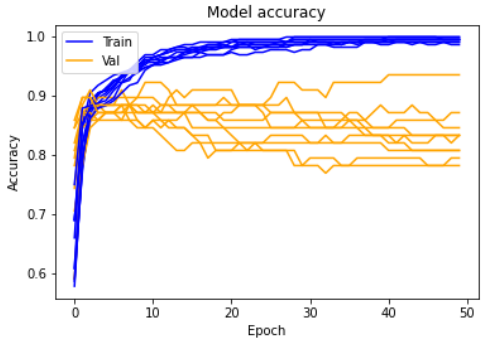

Figure 6B. Accuracy values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 30 minutes.

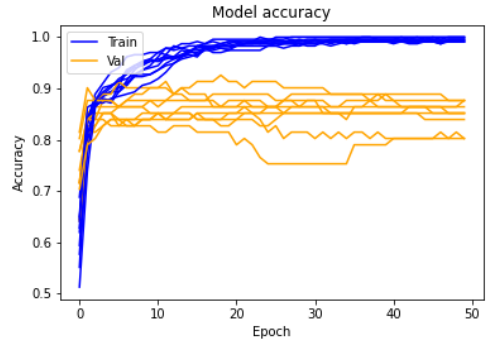

Figure 6C. Accuracy values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 1 hour.

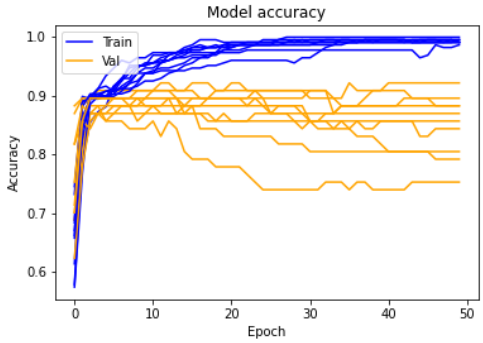

Figure 6D. Accuracy values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 1 hour.

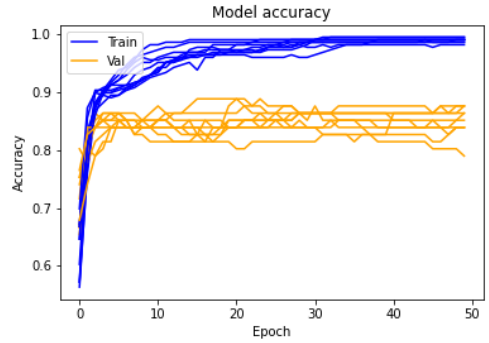

Figure 6E. Accuracy values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 3 hours.

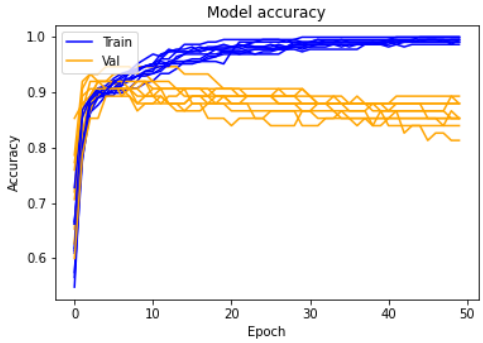

Figure 6F. Accuracy values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 3 hours.

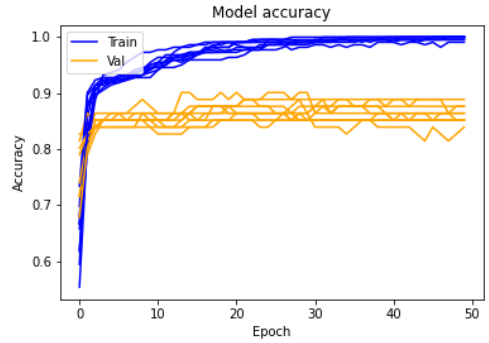

Figure 6G. Accuracy values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 6 hours.

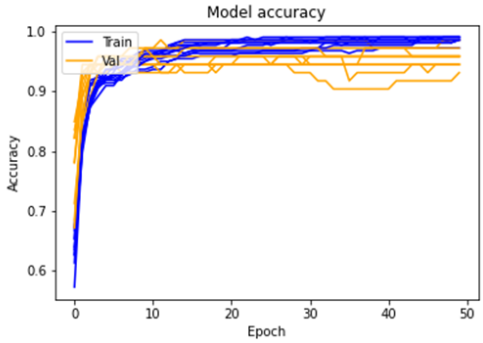

Figure 6H. Accuracy values per epoch during training of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 6 hours.

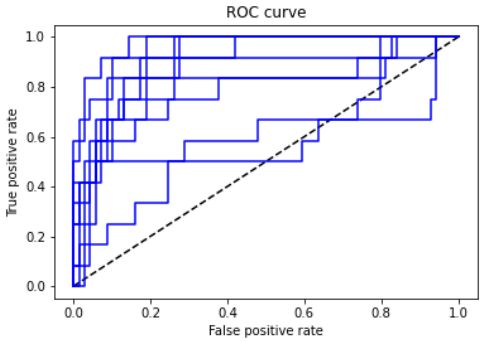

Figure 7A. Receiver Operator Characteristics curve of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes, on test data.

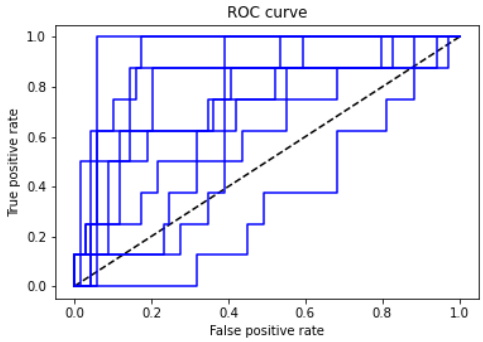

Figure 7B. Receiver Operator Characteristics curve of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 30 minutes, on test data.

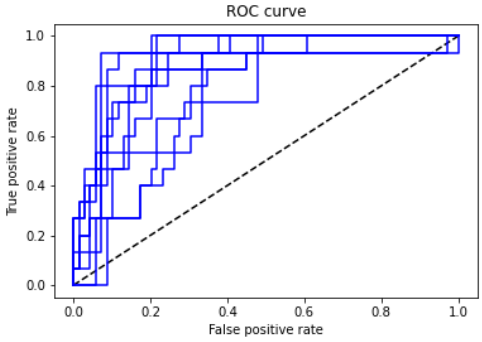

Figure 7C. Receiver Operator Characteristics curve of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 1 hour, on test data.

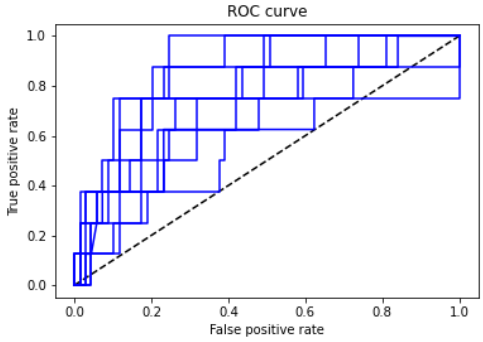

Figure 7D. Receiver Operator Characteristics curve of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 1 hour, on test data.

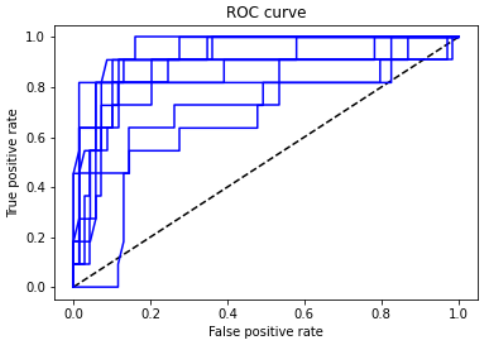

Figure 7E. Receiver Operator Characteristics curve of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 3 hours, on test data.

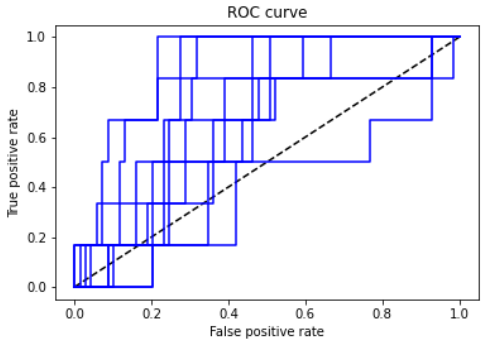

Figure 7F. Receiver Operator Characteristics curve of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 3 hours, on test data.

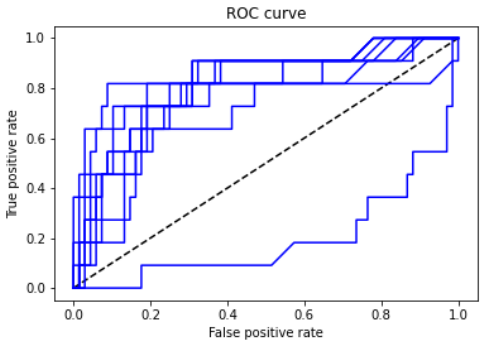

Figure 7G. Receiver Operator Characteristics curve of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 6 hours, on test data.

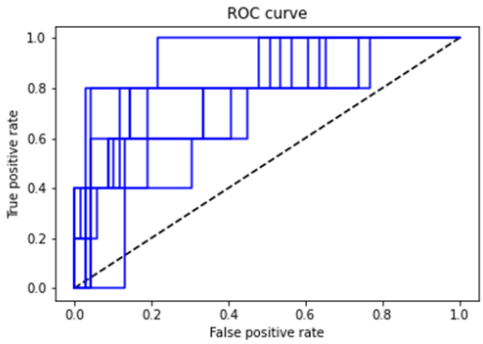

Figure 7H. Receiver Operator Characteristics curve of the 10 models aiming to predict periods of ICP≥20mmHg lasting at least 30 minutes, using a prediction window of 6 hours, on test data.

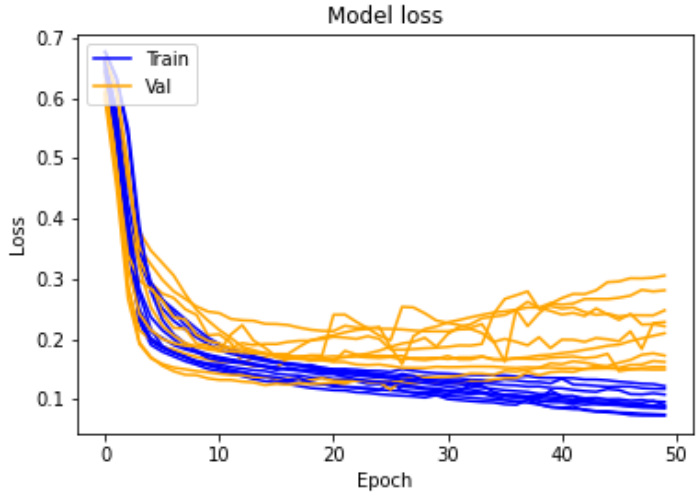

Figure 8. Loss values per epoch during training of the 10 models using only the ICP and MAP, aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes.

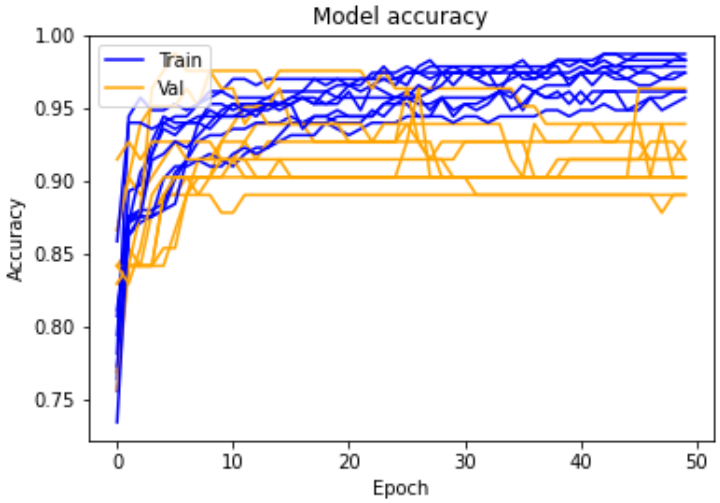

Figure 9. Accuracy values per epoch during training of the 10 models using only the ICP and MAP, aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes.

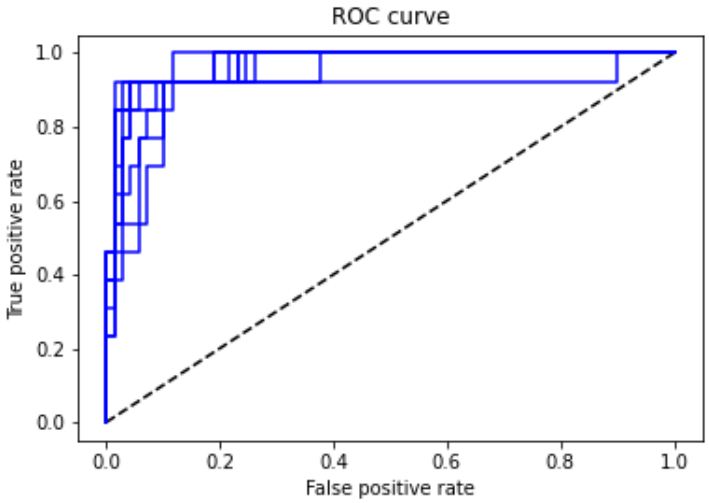

Figure 10. Receiver Operator Characteristics curve of the 10 models using only the ICP and MAP, aiming to predict periods of ICP≥20mmHg lasting at least 15 minutes, using a prediction window of 30 minutes, on test data.